Classifying Out-Of-Vocabulary Terms in a Domain-Specific Social Media Corpus

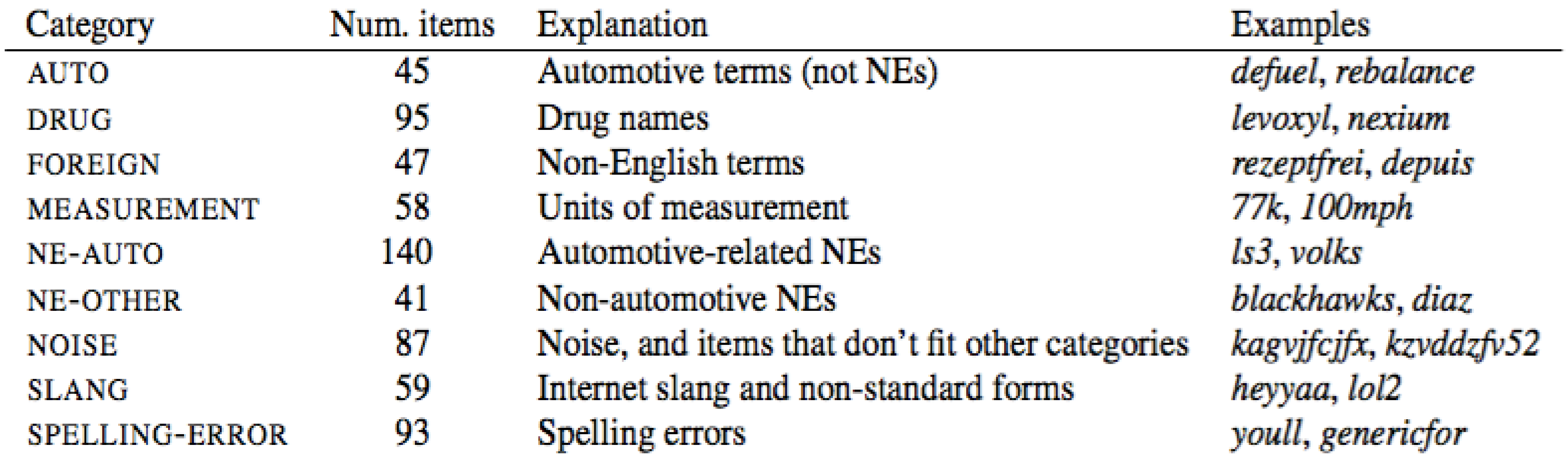
SoHyun Park   Afsaneh Fazly   Annie Lee
Brandon Seibel   Wenjie Zi   Paul Cook

Introduction

- High rate of out-of-vocabulary (OOV) terms in social media text.
- Presents challenge to most natural language processing (NLP) systems as they rely heavily on lexical knowledge.
- Goal: automatically classify OOV terms in automotive web forums into domain specific categories.
- Coarse-grained categories could serve as a preliminary source of lexical knowledge about the out-of-vocabulary terms

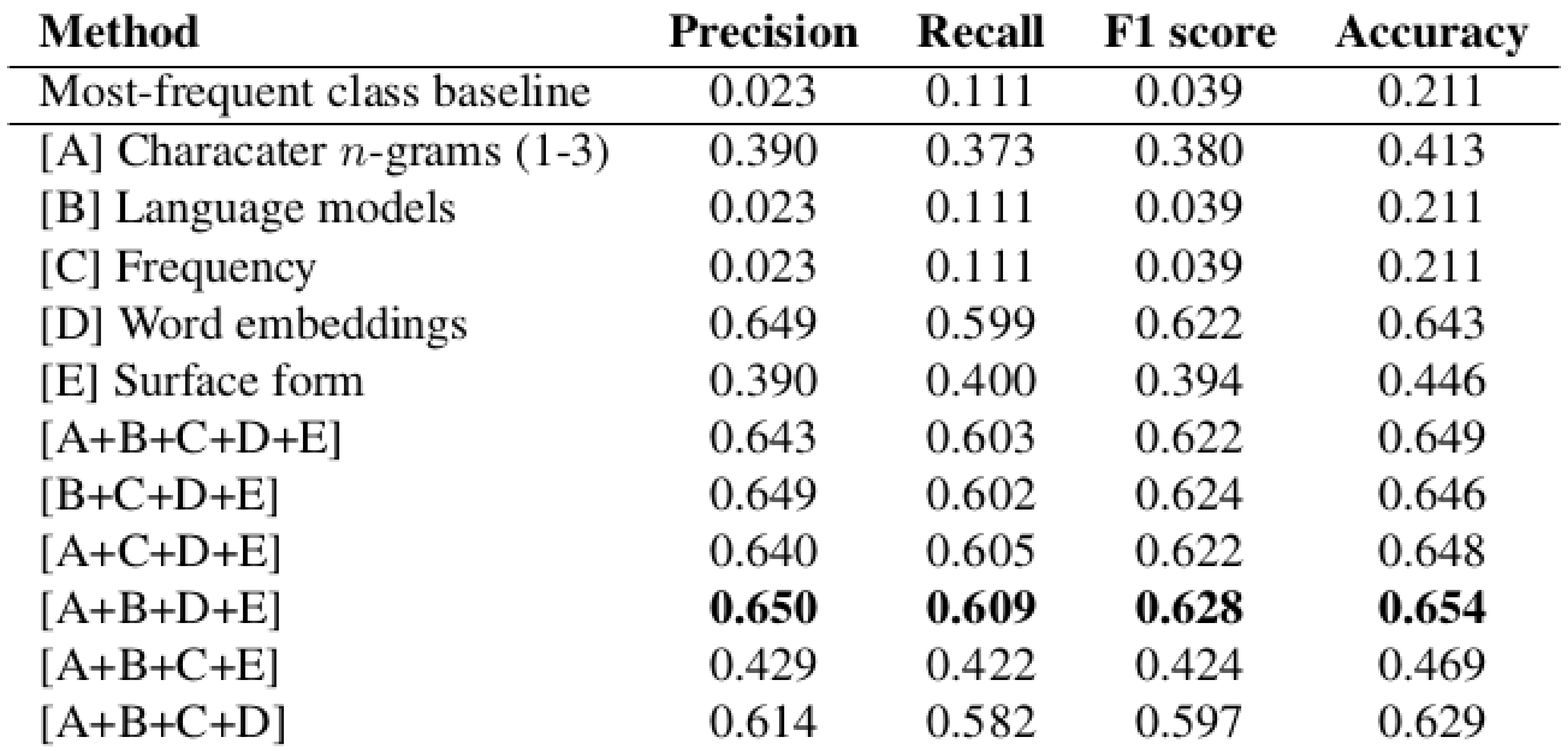
Methods

- Supervised learning approach
- Features sets:
  - Character N-grams
  - Language models
  - Frequency
  - Word embedding
  - Surface form
- Experimental setup: 10x10-fold cross-validation logistic regression



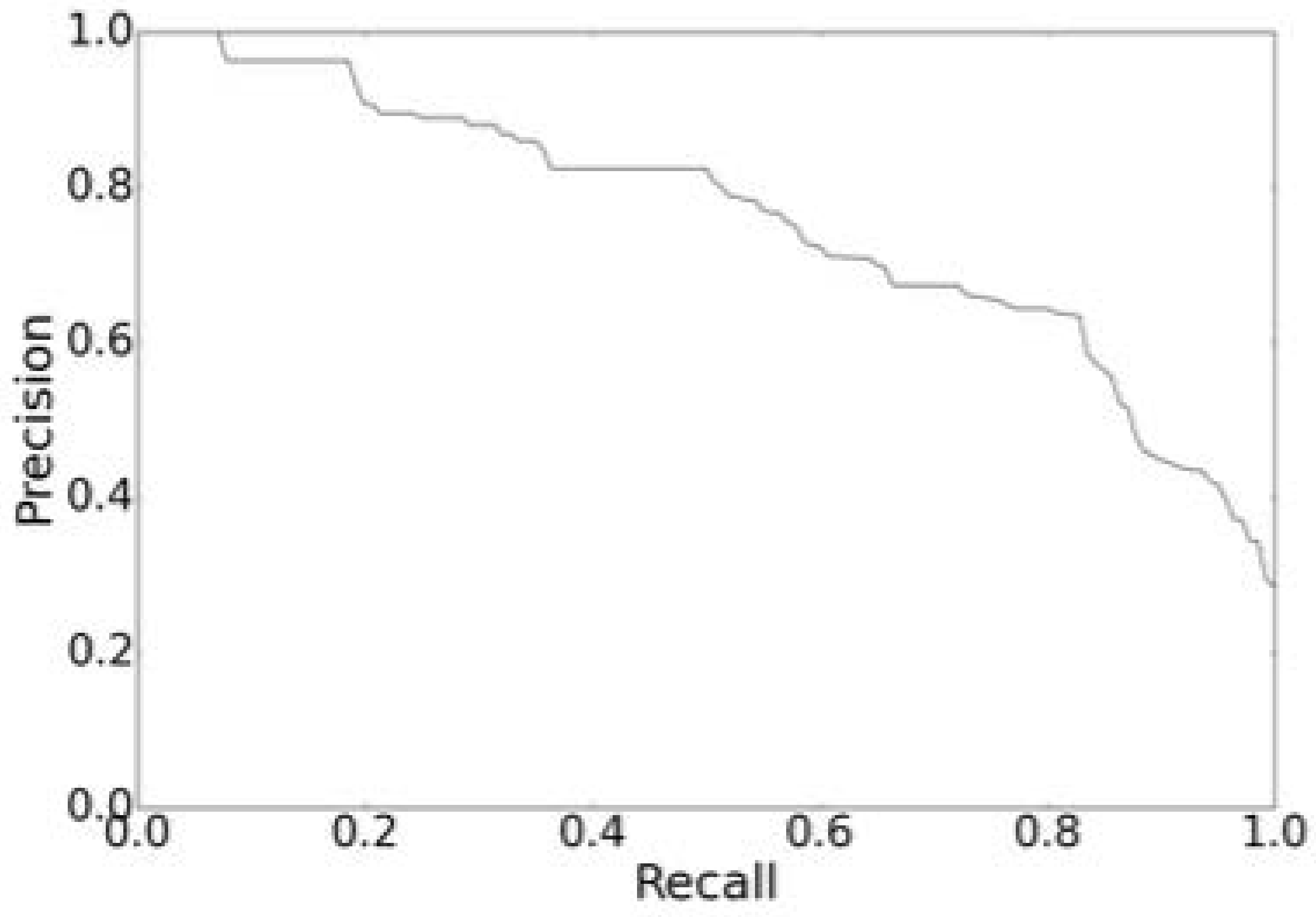
Category	Num. items	Explanation	Examples
AUTO	45	Automotive terms (not NEs)	*defuel, rebalance*
DRUG	95	Drug names	*levoxyl, nexium*
FOREIGN	47	Non-English terms	*rezeptfrei, depuis*
MEASUREMENT	58	Units of measurement	*77k, 100mph*
NE-AUTO	140	Automotive-related NEs	*ls3, volks*
NE-OTHER	41	Non-automotive NEs	*blackhawks, diaz*
NOISE	87	Noise, and items that don't fit other categories	*kagvjfcjfx, kzvddzfv52*
SLANG	59	Internet slang and non-standard forms	*heyyaa, lol2*
SPELLING-ERROR	93	Spelling errors	*youll, genericfor*

Results



Method	Precision	Recall	F1 score	Accuracy
Most-frequent class baseline	0.023	0.111	0.039	0.211
[A] Characater n-grams (1-3)	0.390	0.373	0.380	0.413
[B] Language models	0.023	0.111	0.039	0.211
[C] Frequency	0.023	0.111	0.039	0.211
[D] Word embeddings	0.649	0.599	0.622	0.643
[E] Surface form	0.390	0.400	0.394	0.446
[A+B+C+D+E]	0.643	0.603	0.622	0.649
[B+C+D+E]	0.649	0.602	0.624	0.646
[A+C+D+E]	0.640	0.605	0.622	0.648
[A+B+D+E]	**0.650**	**0.609**	**0.628**	**0.654**
[A+B+C+E]	0.429	0.422	0.424	0.469
[A+B+C+D]	0.614	0.582	0.597	0.629

- Word embedding features are very informative of OOV meaning



- Interpolated precision-recall curve for ranking based on probability of NE-AUTO class
- Ranking can be useful for semi-automatic identification of NE-AUTO terms